Natural Language Processing

Dan Dixey

word frequency counting

Term frequency – inverse document frequency

n-grams

clustering

Word sense disambiguation

Concepts

Data-Driven Event detection:

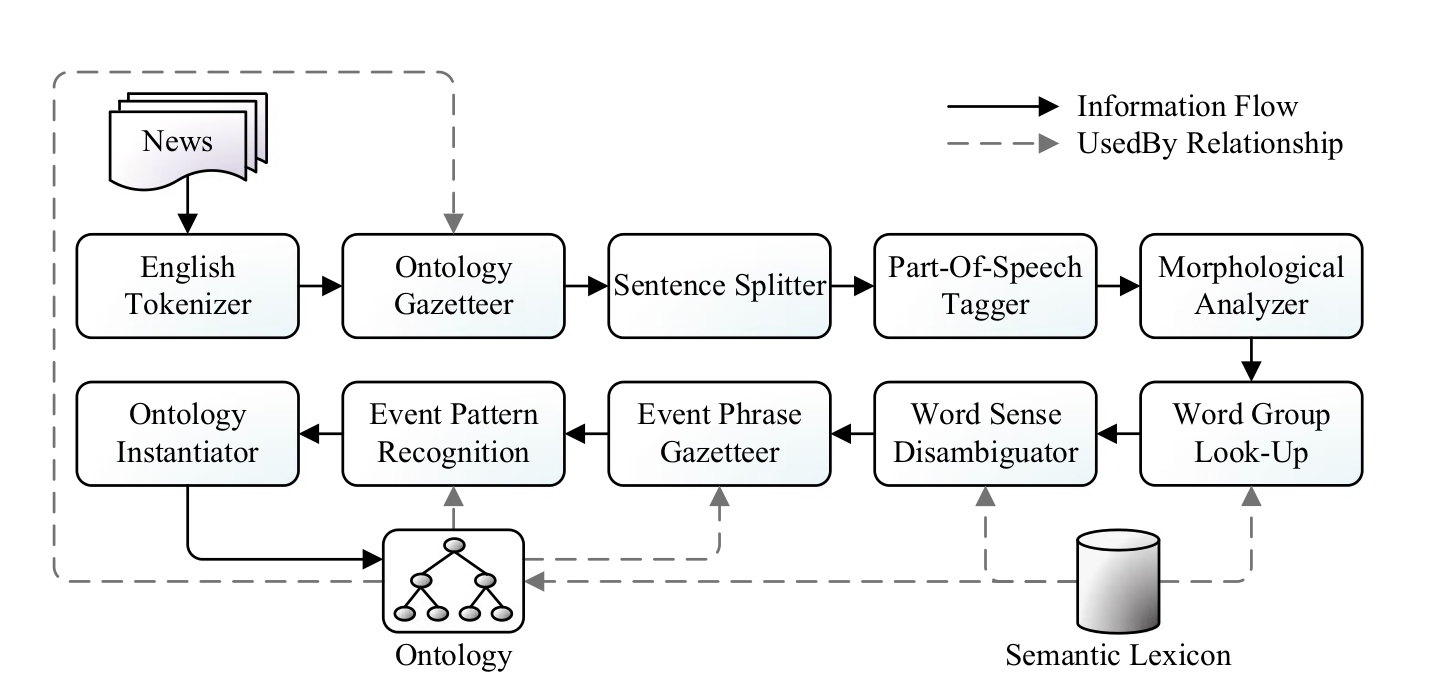
* Okamoto et al. [2009] developed a hierarchical clustering technique for the detection of occasional or local events. While clustering itself could already yield promising results for event extraction the authors of [21] make use of a combination of weighted undirected bipartite graphs and clustering in order to extract key entities and significant events from daily web news.
* Clustering techniques are also employed by Tanev et al. [34], who also aim for real-time news event extraction, but focus especially on violence and disaster events. The authors make use of automatic tagging of words and the presented framework is designed to automatically learn patterns from discovered events.
* Lastly, the authors of [19] also employ word-based statistical text mining in their work from 2005. The authors elaborate on a framework aimed at news event detection, based on support vector machines

Knowledge-Driven Event Extraction:

* In contrast to data-driven methods, knowledge-driven text mining is often based on patterns that express rules representing expert knowledge. It is inherently based on linguistic and lexicographic knowledge, as well as existing human knowledge regarding the contents of the text that is to be processed. This alleviates problems with statistical methods regarding meaning of text. Information is mined from corpora by using predetermined patterns.

Hybrid Driven Event Extraction System:

* We propose a framework (pipeline) that identifies the concepts of interest (i.e., concepts related to economic events), which are defined in a domain ontology and are associated to synsets from a semantic lexicon. In conjunction with statistical techniques for event detection.



Developing a Domain Ontology:

Our envisaged approach is driven by an ontology containing information on the NASDAQ-100 companies, extracted from Yahoo! Finance. This domain ontology has been developed by domain experts through an incremental middle-out approach. The ontology captures concepts and events concerning the financial domain, e.g., companies, competitors, products, CEO’s, etc. Many concepts in this ontology stem from a semantic lexicon (i.e., WordNet) and are linked to their semantic lexicon counterparts, but a significant part of the ontology consists of concepts representing named entities (i.e., proper names). In our ontology, we distinguish between ten different financial events, i.e., announcements regarding CEOs, presidents, products, ompetitors, partners, subsidiaries, share values, revenues, profits, and losses, which are supported by appropriate classes and properties. We validated our domain ontology using OntoClean [15], a methodology for analysing ontologies that uses notions for philosophical ontological analysis. OntoClean is based on formal, domain-independent class properties (meta-properties and their codifiers), i.e., identity, unity, rigidity, and dependence. Once annotated with these meta-properties, the ontology can be considered to be valid (or “clean”) whenever no constraints are violated that are based on these properties.

Ontology Gazetter:

A first step towards understanding the text is subsequently taken by the Ontology Gazetteer, which links concepts in the text to concepts defined in an ontology with relevant concepts (which tend to refer to proper names rather than common words from the semantic lexicon). A normal gazetteer uses lists of words as input, whereas our ontology gazetteer is ontology-driven and scans the text for lexical representations of concepts from the ontology. Matching tokens in the text are annotated with a reference to their associated concepts defined in the ontology. For example, suppose our ontology contains a concept ‘Google’ of type ‘Company’, with a lexical representation ‘Google Inc.’. Any matching ‘Google Inc.’ occurrence in the text is then annotated with the concept ‘Google’. The default GATE OntoGazetteer uses a linear search algorithm to match lexical representations in a text with a list of ontology concepts and their associated lexical representations. However, in our novel OntoLookup approach, we use a look-up tree of approximately 5,000 nodes (based on the Yahoo! Finance news messages represented in the ontology), in which possible lexical representations of all relevant concepts in the ontology are mapped to their associated concepts. Each concept can have multiple lexical representations (groups of 1 or more words). These word groups are all represented in the look-up tree. Nodes in the tree represent individual tokens and a path from the root node to an arbitrary leaf node represents a word group.

Event Phrase Gazetteer:

When the meaning of word groups has been disambiguated, the text can be in interpreted using semantics introduced by linking word groups to an ontology, thus capturing their essence in a meaningful and machine-understandable way. As we are interested in specific economic events, the Event Phrase Gazetteer scans the text for those events. It uses a list of phrases or concepts that are likely to represent some part of a relevant event. For ample, when we are looking for stock splits, we can search for ‘stock split’. Since the Word Group Look-Up component has already combined ‘stock’ and ‘split’ and the Word Sense Disambiguator has already assigned a concept value to this group of words, we can easily match this concept with events in our ontology.

References:

1. Borsje J, Hogenboom F, Frasincar F (2010) Semi-automatic financial events discovery based on exico-semantic patterns. Int J Web Eng Technol 6(2):115–140
2. Bechhofer S, van Harmelen F, Hendler J, Horrocks I, McGuinness D, Patel-Schneider P, Stein (2004) OWL Web ontology language reference. From: <http://www.w3.org/TR/owl-ref/>

Links:

<http://ontotext.com/products/ontotext-graphdb/>

<http://dl.acm.org/citation.cfm?id=1699766>-A novel approach to automatic gazetteer generation using Wikipedia